Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2008 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

July 2008

Collaborative Filtering-based Context-Aware Document-Clustering (CF-CAC) Technique

Chih-Ping Wei
National Tsing Hua University, cpwei@mx.nthu.edu.tw

Chin-Sheng Yang
National Sun Yat-sen University, litony@gmail.com

Collaborative Filtering-based Context-Aware Document-Clustering (CF-CAC) Technique

Chin-Sheng Yang, Department of Information Management, College of Management, National Sun Yat-sen University, Kaohsiung, Taiwan, R.O.C., litony@gmail.com
Chih-Ping Wei, Institute of Technology Management, College of Technology Management, National Tsing Hua University, Hsinchu, Taiwan, R.O.C., cpwei@mx.nthu.edu.tw

Abstract
Document clustering is an intentional act that should reflect an individual's preference with regard to the semantic coherency or relevant categorization of documents and should conform to the context of a target task under investigation. Thus, effective document clustering techniques need to take into account a user's categorization context. In response, Yang & Wei (2007) propose a Context-Aware document Clustering (CAC) technique that takes into consideration a user's categorization preference relevant to the context of a target task and subsequently generates a set of document clusters from this specific contextual perspective. However, the CAC technique encounters the problem of small-sized anchoring terms. To overcome this shortcoming, we extend the CAC technique and propose a Collaborative Filtering-based Context-Aware document-Clustering (CF-CAC) technique that considers not only a target user's but also other users' anchoring terms when approximating the categorization context of the target user. Our empirical evaluation results suggest that our proposed CF-CAC technique outperforms the CAC technique.

Keywords: *Document clustering, Context-aware document-clustering, Collaborative filtering, Text mining*

1. Introduction
With the advances and proliferation of the Internet, available information sources have grown tremendously in number and sheer volume, primarily as a result of global connectivity and ease of publishing. To manage this ever-increasing volume of documents, organizations and individuals typically organize documents into categories (or category hierarchies) to facilitate their document management and to support subsequent document retrieval and access. Hence, the development of an effective document clustering mechanism becomes essential to efficient and effective document management of organizations and individuals.

Document clustering entails the automatic organization of a large document collection into distinct groups of similar documents that reflect general themes hidden within the corpus (Pantel & Lin, 2002; Wei et al., 2006b). However, according to the context theory of classification, document clustering behaviors of individuals not only involve the attributes (including contents) of documents but also depend on who is performing the task and in what context (Barreau, 1991; Case, 1991; Kwasnik, 1991; Lakoff, 1987). As a result, document clustering is an intentional act that should reflect individuals' preferences with regard to the semantic coherency or relevant categorization of documents (Rucker & Polanco, 1997) and should conform to the context of a target task under investigation. For example, given a set of research articles related to "data mining," an individual who is interested in developing new data mining techniques may prefer a set of document categories anchored at techniques under discussion, whereas the same individual may prefer a different set of document categories based on application domains involved when he/she is working on data mining applications. The aforementioned examples highlight the importance of clustering the same set of documents into different document categories for different task contexts concerned by the same individual. Effective document clustering techniques therefore need to be able to take into account a user's categorization context defined by or relevant to the target task under consideration.

Traditional document clustering techniques generally anchor in pure content-based analysis. That is, most of existing document clustering techniques rely on a specific feature selection metric (e.g., term frequency (TF) or TF×IDF (term frequency×inverse document frequency)) (Boley et al., 1999; Larsen & Aone, 1999; Pantel & Lin, 2002; Roussinov & Chen, 1999; Wei et al., 2006b) that are objective in nature to identify a set of representative features as the basis for document clustering. Consequently, existing document clustering techniques create a set of clusters that are not tailored to individuals' categorization contexts and therefore are not able to facilitate context-aware document-clustering. The categorization scheme exhibited in such context-unaware clusters may not conform to that of an individual's expectations and perceptions under a specific context. However, an individual's document search typically is guided by his/her own categorization scheme (Donovan, 1991; Restorick, 1986). Thus, when searching documents with a one-for-all categorization scheme, an individual generally undertakes a semantic internalization process (Quillian, 1968) to comprehend the target categorization scheme or experiences a coadaptation process that adjusts his/her own categorization scheme and, at the same time, reinterprets and adapts the target categorization scheme to his/her needs (Mackay, 1988; Mackay, 2000). The semantic internalization and coadaptation processes unnecessarily increase the individual's cognitive load. As a result, he/she likely spends more time or has difficulty locating documents of interest because of the discrepancy between the one-for-all categorization scheme and his/her expectation (Wei et al., 2006a). The described inefficiency or ineffectiveness of document retrieval and access may adversely affect the efficiency, quality, and satisfaction of decision making that requires references to various documents relevant to the target decision context.

In response to the limitations of existing document clustering techniques and the needs of supporting context-aware document-clustering, Yang & Wei (2007) propose a Context-Aware document-Clustering (CAC) technique that takes into consideration a user's categorization preference (expressed as a list of anchoring terms) relevant to the context of a target task and subsequently generates a set of document clusters from this specific contextual perspective. However, the effectiveness of the CAC technique is sensitive to the size of anchoring terms. That is, as the size of anchoring terms decreases, the effectiveness of the CAC technique deteriorates sharply. To overcome this shortcoming, we extend the CAC technique and propose a Collaborative Filtering-based Context-Aware document-Clustering (CF-CAC) technique that considers not only a target user's but also other users' anchoring terms when approximating the categorization context of the target user. Specifically, adopting the collaborative filtering recommendation concept, the CF-CAC technique first determines a set of neighbors whose categorization contexts are similar to that of the target user and then expands the target user's categorization context (i.e., anchoring terms) by considering those of his/her neighbors. The expanded categorization context of the target user subsequently becomes the input to the CAC technique.

The remainder of this paper is organized as follows. Section 2 reviews existing document clustering techniques relevant to this study. In Section 3, we depict the detailed design of the proposed CF-CAC technique. Subsequently, we describe our experimental design and discuss important evaluation results in Section 4. Finally, we conclude in Section 5with a summary and some future research directions.

2. Literature Review

2.1 Content-based Document Clustering Techniques

In essence, document clustering groups similar documents into clusters. The documents in the resultant clusters exhibit maximal similarity to those in the same cluster and, at the same time, share minimal similarity with documents in other clusters. Most of existing document clustering techniques are anchored in document content analysis. The overall process of a content-based document clustering technique generally comprises three main phases: feature extraction and selection, document representation, and clustering (Jain et al., 1999; Wei et al., 2002; Wei et al., 2006b). The purpose of feature extraction and selection is to extract and select from the target document corpus a set of representative features to represent the documents in the document representation phase. Subsequently, the clustering phase applies a clustering technique to group the target documents into

distinct clusters.

Feature extraction begins with the parsing of each source document to produce a set of nouns and noun phrases and exclude a list of prespecified "stop words" that are non-semantic-bearing words. Subsequently, representative features are selected from the set of extracted features. Feature selection is important for clustering efficiency and effectiveness, because it not only condenses the size of the extracted feature set, but also reduces the potential biases embedded in the original (i.e., nontrimmed) feature set (Roussinov & Chen, 1999; Yang & Chute, 1994). Commonly used feature selection metrics include: TF, TF×IDF, and their hybrids (Boley et al., 1999; Larsen & Aone, 1999).

On the basis of a particular feature selection metric, the k features with the highest selection metric scores then are selected to represent each source document in the document representation phase. Based on the chosen representation scheme, each document is described in the k-dimensional space and represented as a feature vector. Commonly employed document representation schemes include binary (presence or absence of a feature in a document), within-document TF, and TF×IDF (Boley et al., 1999; Larsen & Aone, 1999; Pantel & Lin, 2002; Roussinov & Chen, 1999; Wei et al., 2006b).

In the final phase of document clustering, source documents are grouped into distinct clusters on the basis of the selected features and their respective values in each document. Common clustering approaches include partitioning-based (Boley et al., 1999; Cutting et al., 1992; Larsen & Aone, 1999), hierarchical (El-Hamdouchi & Willett, 1986; Roussinov & Chen, 1999; Voorhees, 1986; Wei et al., 2006b), and Kohonen neural network (Lagus et al., 1996; Lin et al., 1999-2000; Roussinov & Chen, 1999).

As mentioned, content-based document clustering techniques rely on an objective feature-selection metric (e.g., TF or TF×IDF) that merely considers document content. As a result, existing content-based techniques generate for all users an identical set of document clusters from a given document collection and, thus, is unable to support context-aware document-clustering.

2.2 Context-Aware Document-Clustering (CAC) Technique

In response to the shortcomings and limitations of existing document clustering techniques for supporting context-aware document-clustering, Yang and Wei (2007) propose a Context-Aware document-Clustering (CAC) technique that takes into consideration a user's categorization preference (expressed as a list of anchoring terms) relevant to the context of a target task and subsequently generates a set of document clusters from this specific contextual perspective. The CAC technique consists of five main phases: 1) feature extraction and selection; 2) statistical-based thesaurus construction; 3) anchoring term expansion; 4) document representation; and 5) clustering.

The feature extraction and selection aims at extracting and selecting a set of representative features from the target document corpus. Furthermore, features that infrequently appear in the target document corpus are removed. Particularly, only those features whose document frequency is no less than a prespecified threshold δ_{DF} are retained. This set of representative features forms the basis for anchoring term expansion.

The purpose of the statistical-based thesaurus construction phase is to automatically construct a statistical-based thesaurus that will be used for expanding the user-provided anchoring terms. CAC exploits the World Wide Web (WWW) to create the statistical-based thesaurus, which will serve as the basis for expanding the set of anchoring terms relevant to the categorization context of a user.

For each anchoring term q_i pertaining to the categorization context of a user and every feature f_j representative to the target document corpus, CAC issues three queries (i.e., q_i, f_j, and $q_i \wedge f_j$) to a search engine and obtains the number of hits (matched documents) returned for each query. The

relevance weight between q_i and f_j is then estimated by the pointwise mutual information (PMI) (Turney & Littman, 2003) as follows:

$$rw_{q_i f_j} = \log_2 \left(\frac{p(q_i \wedge f_j)}{p(q_i)\, p(f_j)} \right) = \log_2 \left(\frac{N \times \text{hits}(q_i \wedge f_j)}{\text{hits}(q_i)\, \text{hits}(f_j)} \right)$$

where $rw_{q_i f_j}$ denotes the relevance weight between q_i and f_j, $p(query)$ is the probability that *query* occurs in the repository (i.e., WWW in their study), N is total number of documents in the repository, and hits(*query*) is the number of hits returned by the search engine of choice.

On the basis of the statistical-based thesaurus constructed, the expansion of anchoring term is to expand the set of anchoring terms AT by including additional relevant terms. An anchoring term q_i in AT is expanded with a set of terms E_{q_i} whose relevance weights to q_i need to be greater than a prespecified threshold α. Accordingly, the resultant expanded set of anchoring terms $RF = \left(\bigcup_{q_i \in AT} E_{q_i} \right) \cup AT$ is formed for the subsequent document clustering task.

Because RF consists of the anchoring terms originally provided by the user and relevant terms expanded from the anchoring terms, the importance of the terms in RF should not be identical when they are used to represent each document to be clustered. Accordingly, CAC adopts the TF×IDF-like scheme and defines the weight of each expanded term f_j in RF but not in AT as:

$$w_j = \sum_{q_i \in ET_j} rw_{q_i f_j} \times \log \left(\frac{|AT|}{|ET_j|} + \varepsilon \right)$$

where ET_j is the set of anchoring terms that expand f_j and ε is a small positive value to avoid the log component in the formula being 0. On the other hand, if $f_i \in AT$, w_j is the largest weight across all expanded terms derived previously.

In the document representation phase, each document to be clustered is represented using the expanded set of anchoring terms RF. CAC employs the TF×IDF scheme weighted by the weight of each term in the expanded set of anchoring terms for document representation. Finally, in the clustering phase, the target documents are grouped into distinct clusters on the basis of the expanded set of anchoring terms (i.e., RF) and their respective values in each document. CAC adopts the hierarchical clustering approach (specifically, the HAC algorithm) as the underlying clustering algorithm.

Though the effectiveness of the CAC technique is encouraging, it is susceptible to the size of the anchoring terms. However, in a typical real-world setting, the set of anchoring terms provided by a user often tends to be small; therefore, the CAC technique needs to be enhanced so that it can effectively cluster documents even when only a small-sized set of anchoring terms is available.

3. Collaborative Filtering-based Context-Aware Document-Clustering (CF-CAC) Technique

We propose the CF-CAC technique in response to the abovementioned limitations of the CAC technique in the situation where only a small-sized set of anchoring terms that partially describes a user's categorization context is available. In this study, the CF-CAC technique considers not only the target user's but also other users' anchoring terms when approximating the categorization context of the target user. Specifically, adopting the collaborative filtering recommendation concept, the CF-CAC technique first determines a set of neighbors whose categorization contexts are similar to that of the target user and then expands the target user's categorization context (i.e., anchoring terms) by considering those of his/her neighbors. Subsequently, the expanded categorization context of the target user becomes the input to the existing CAC technique. As Figure 1 illustrates, the overall process of the CF-CAC technique consists of five phases: 1) collaborative context expansion; 2) feature extraction and selection; 3) anchoring term expansion; 4) document representation; and 5) clustering.

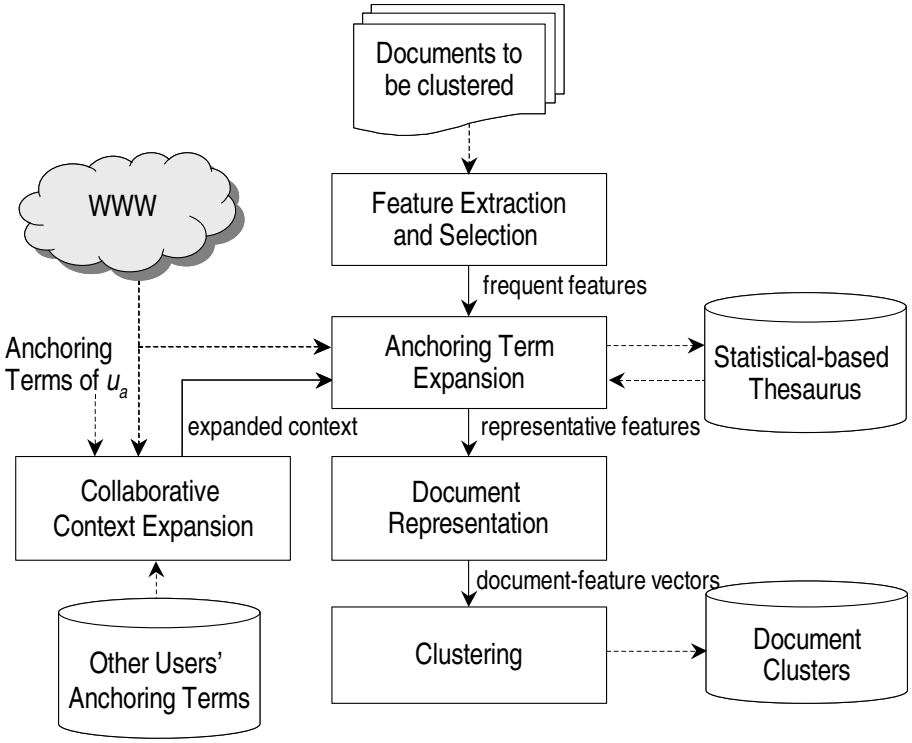

Figure 1: Overall Process of the CF-CAC Technique

Collaborative Context Expansion: This phase aims at expanding the target user's categorization context by taking into consideration the target user's anchoring terms and those of other users with similar categorization contexts. Two major tasks are involved in this phase: neighborhood formation and context expansion.

To form the neighborhood for a specific target user u_a, we estimate the similarities between the target user and all other users on the basis of their anchoring terms. The simplest and intuitive method is to treat the anchoring terms of two users as two sets and then compute the similarity of these two sets by employing a similarity measure such as Jaccard or Dice. However, several problems possibly limit the practicability of the abovementioned method. For example, because the anchoring terms are specified by a user according to his/her preferences and perception (i.e., without any hints or additional supports), it is difficult to guarantee that two users will use an identical term for describing the same concept. Some linguistic variations, such as orthographic variations, morphological variations, abbreviation, and acronym, of the anchoring terms increase the difficulty of the target similarity estimation. Even when most of the linguistic variations can be addressed via some appropriate text processing mechanisms (e.g., stemming, linguistic dictionary checking, etc.); there still exists another challenge, specifically the word mismatch problem, to the aforementioned similarity estimation method. Word mismatch refers to the phenomenon in which people use different terms to describe the same concept. According to Furnas et al.'s (1987) study, the probability that two people will use an identical term or terms to describe the same concept (or object) is less than 20%. For example, some people might use the term "data mining" to describe the process or techniques for extracting novel, valid, and actionable patterns from databases, whereas others may choose "knowledge discovery" or "data archeology" to refer to the same concept.

In response, we propose an alternative context similarity estimation method that employs World Wide Web (WWW) as the information source to estimate the similarity between two sets of anchoring terms. Assume that $q_i \in AT_a$ is an anchoring term of user u_a and $q_j \in AT_b$ is an anchoring term of another user u_b. First, for each pair of anchoring terms q_i and q_j of the two users, we issue three queries (i.e., q_i, q_j,

and $q_i \wedge q_j$) to a search engine (particularly, Google in this study) and obtain the number of hits (matching documents) returned for each query. We then estimate the relevance weight rw_{q_i,q_j} between a pair of anchoring terms q_i and q_j by the pointwise mutual information (PMI) measure (Turney & Littman, 2003) as follows:

$$rw_{q_i,q_j} = \log_2 \left(\frac{p(q_i \wedge q_j)}{p(q_i)\, p(q_j)} \right) = \log_2 \left(\frac{N \times \text{hits}(q_i \wedge q_j)}{\text{hits}(q_i)\, \text{hits}(q_j)} \right),$$

where $p(query)$ is the probability that *query* occurs in the repository (i.e., WWW in this study), N is the total number of documents in the repository, and hits(*query*) is the number of hits returned by the search engine of choice. Because the exact value of N in the WWW environment is difficult to estimate, we set N as the largest hit value among all the queries issued to the search engine.

A prespecified threshold λ is applied to remove insignificant relevance weights. In other words, a pair of terms whose rw_{q_i,q_j} is no less than λ is considered as related terms. Subsequently, we standardize all relevance weights between all pairs of terms to 0 to 1. After the estimation and standardization of the relevance weight of each pair of anchoring terms, we estimate the similarity from the set of anchoring terms of the target user u_a (denoted AT_a) to that of another user u_b (denoted AT_b). The context similarity from u_a to u_b is then computed as follows:

$$Similarity(u_a \rightarrow u_b) = Similarity(AT_a, AT_b) = \frac{1}{|AT_a|} \sum_{q_i \in AT_a} sim(q_i, AT_b)$$

$$\text{where } sim(q_i, AT_b) = \begin{cases} 1 & \text{if } q_i \in AT_b \\ \underset{q_j \in AT_b \text{ and } q_i \text{ and } q_j \text{ are related terms}}{\text{average}}(rw_{q_i,q_j}) & \text{otherwise} \end{cases}.$$

After we compute the context similarities from the target user u_a to all other users, we select a set of candidate neighbors CN_a with top-ranked $Similarity(u_a \rightarrow u_b)$. This candidate neighbor selection process ensures that the categorization contexts of the users in CN_a are similar to that of the target user u_a.

Furthermore, for each $u_b \in CN_a$, we transform the $Similarity(u_b \rightarrow u_a)$ into an importance score by the following exponential equation. This transformation process attempts to ensure that those users with higher importance scores not only are similar to the target user u_a in categorization context but also have the potential for expanding additional anchoring terms. We then form the neighborhood N_a for u_a by selecting the top n most important users.

$$Importance(u_b \rightarrow u_a) = \exp(-|Similarity(u_b \rightarrow u_a) - 0.5|).$$

After the neighborhood formation task, the context expansion task is undertaken to address the problem of a possibly small-sized set of anchoring terms of u_a that degrades the effectiveness of the CAC technique. Specifically, the expanded context (i.e., an expanded set of anchoring terms) $EAT_a = \left(\bigcup_{u_b \in N_a} AT_b \right) \cup AT_a$ for the target user u_a is the union of u_a's and all his/her neighbors' anchoring terms. For each anchoring term q_j in EAT_a but not in AT_a, we estimate its weight by summing up the $Similarity(u_a \rightarrow u_b)$ of those users that expand q_j. That is $ew_{qj} = \sum_{u_b \in N_a} Similarity(u_a \rightarrow u_b)$. On the other hand, for those anchoring terms originally pertain to AT_a, their weights are the largest weight across all expanded anchoring terms derived previously. Finally, we standardize the weights of all anchoring terms in EAT_a to the interval of 0 to 1.

Feature Extraction and Selection: The purpose of this phase is to extract and select a set of representative features (specifically, nouns and noun phrases) from the target document corpus (i.e., the collection of documents to be clustered). This set of representative features forms the basis for anchoring term expansion. We adopt the rule-based part-of-speech tagger developed by Brill (1994) to

syntactically tag each word in the target documents. Subsequently, this study employs the approach proposed by Voutilainen (1993) to implement a noun-phrase parser for extracting noun phrases from each syntactically tagged document. Furthermore, we remove features that infrequently appear in the target document corpus. Particularly, we only retain those features whose document frequency is no less than a prespecified threshold δ_{DF}.

Anchoring Term Expansion: The purpose of this phase is to expand the set of anchoring terms in the expanded context by including additional relevant terms. Specifically, two major tasks are performed in this phase, namely statistical-based thesaurus construction and expansion of anchoring terms. The purpose of statistical-based thesaurus construction is to automatically construct a statistical-based thesaurus that will be used for expanding the anchoring terms EAT_a relevant to the target user u_a's expanded categorization context. As with the CAC technique, we exploit the World Wide Web (WWW) to create the statistical-based thesaurus, because WWW probably is the largest repository in the world and the association strength (or relevance weight) between two terms measured by the co-occurrence analysis on a search engine's query results will have higher statistical reliability than that estimated from the co-occurrence analysis on a smaller document corpus (Turney & Littman, 2003).

For each anchoring term q_i pertaining to the expanded categorization context EAT_a of the target user u_a and a feature f_j representative to the target document corpus, we issue three queries (i.e., q_i, f_j, and $q_i \wedge f_j$) to a search engine (specifically, Google is adopted in this study) and obtain the number of hits (matching documents) returned for each query. We denote the set of queries for the intended clustering task for the target user u_a as a context-aware document-clustering session. The relevance weight between q_i and f_j is then estimated by the pointwise mutual information (PMI) (Turney & Littman, 2003) as follows:

$$rw_{q_i f_j} = \log_2 \left(\frac{p(q_i \wedge f_j)}{p(q_i)\, p(f_j)} \right) = \log_2 \left(\frac{N \times \text{hits}(q_i \wedge f_j)}{\text{hits}(q_i)\, \text{hits}(f_j)} \right)$$

where $rw_{q_i f_j}$ denotes the relevance weight between q_i to f_j, $p(query)$ is the probability that $query$ occurs in the repository (i.e., WWW), N is the total number of documents in the repository, and hits($query$) is the number of hits returned by the search engine of choice. Because the exact value of N in the WWW environment is difficult to estimate, we set N as the largest hit value among all the queries issued in the target context-aware document-clustering session for the user u_a.

With the use of the statistical-based thesaurus constructed, the expansion of anchoring terms is to expand the set of anchoring terms EAT_a encompassed in the expanded categorization context of u_a by including additional relevant terms. Specifically, an anchoring term q_i in EAT_a is expanded with a set of terms E_{q_i} whose relevance weights to q_i need to be greater than a prespecified threshold α. Accordingly, the resultant expanded set of anchoring terms $RF_a = \left(\bigcup_{q_i \in EAT_a} E_{q_i} \right) \cup EAT_a$ is constructed for the subsequent document clustering task.

Because RF_a consists of the anchoring terms provided originally by the target user u_a, expanded in the collaborative context-expansion phase, and expanded in this phase, the importance of these terms in RF_a should not be identical when they are used to represent each document to be clustered. Accordingly, we adopt the TF×IDF-like scheme and define the weight of each expanded term f_j in RF_a but not in EAT_a as:

$$w_j = \sum_{q_i \in ET_j} (ew_{q_i} \times rw_{q_i f_j}) \times \log \left(\frac{|EAT_a|}{|ET_j|} + \varepsilon \right)$$

where ET_j is the set of anchoring terms that expand f_j and ε is a small positive value to avoid the log component in the formula being 0.

On the other hand, if $f_i \in AT_a$, w_j is the largest weight (i.e., $w_{\max}$) across all expanded terms derived

previously. Finally, if $f_i \in EAT_a$ but $f_i \notin AT_a$, w_j is the larger value of ($w_{\max} \times ew_j$) and $\sum_{q_i \in ET_j} rw_{q_i,f_j} \times \log\left(\frac{|EAT_a|}{|ET_j|} + \varepsilon\right)$.

Document Representation: This phase represents each document to be clustered using the expanded set of anchoring terms RF_a. In this study, we employ the TF×IDF scheme weighted by the weight of each term in the expanded set of anchoring terms for document representation. Specifically, each document d_l is described by a feature vector $\vec{d_l}$ as:

$$\vec{d_l} = <v_{l1} \times w_1, v_{l2} \times w_2, \ldots, v_{lm} \times w_m>,$$

where m is the total number of terms in RF_a, v_{lj} is the TF×IDF value of f_j in d_l, and w_j is the weight of term f_j in RF_a.

Clustering: In the final phase, the target documents are grouped into distinct clusters on the basis of the expanded set of anchoring terms (i.e., RF_a) and their respective values in each document. Among the common document clustering approaches (including partitioning-based, hierarchical, and Kohonen neural network), hierarchical clustering has an advantage over partitioning-based, in that the number of clusters need not be prespecified and can be decreased (or increased) by adjusting the intercluster similarity threshold. Furthermore, the hierarchical clustering approach could achieve clustering effectiveness comparable to the Kohonen neural network (Roussinov & Chen, 1999). Therefore, we adopt the hierarchical clustering approach (specifically, HAC) as the underlying clustering algorithm for our proposed CF-CAC technique. In addition, we adopt the cosine measure to estimate the similarity between two documents and employ the group-average link method for measuring the similarity between two clusters.

4. Empirical Evaluation

4.1 Data Collection

To evaluate the effectiveness of the CF-CAC technique, we require three types of data collection, including a document corpus, individuals' preferential categorization contexts and their preferred clusterings for the document corpus, and the categorization contexts of other users serving as possible neighbors. The collection of document corpus for our evaluation purpose consists of 434 research articles related to information systems and technologies that are collected through keyword searches (e.g., XML, data mining, robotics) from a scientific literature digital library website (i.e., CiteSeer, http://citeseer.nj.nec.com/). For each article in our CiteSeer corpus, only the abstract and keywords are used in this evaluation study.

We develop a Web-based system to collect individuals' preferential categorization contexts and their preferred clusterings for the CiteSeer corpus. Each experimental subject is asked to categorize the randomly ordered documents manually. After clustering, the subject is asked to assign a label for each category. These category labels are then considered as the set of anchoring terms with respect to the categorization context relevant to his/her clustering of the corpus and will be used as the input to the CF-CAC technique. A total of 33 subjects accomplish the manual clustering of the documents in the CiteSeer Corpus. According to the self-reported estimates of the subjects, each subject spends a minimum of eight hours performing manual document clustering. A summary of the document categories generated by the subjects is provided in Table 1. Furthermore, we estimate the intersubject agreement of complete clustering results and user-provided anchoring terms among these 33 subjects using the Jaccard and Dice similarities. The average Jaccard and Dice similarities of complete clustering results among the 33 subjects are 33.53% and 49.25% respectively, while the average Jaccard and Dice similarities of anchoring terms among these subjects are 20.84% and 33.82% respectively.

Table 1: Summary of Subjects' Clusterings for the CiteSeer Corpus

	Number of Folders*	Number of Documents in a Folder
Maximum	67	125
Minimum	10	1
Average	26.12	16.64

*: Number of folders equals to the number of anchoring terms specified by a target subject in our experiments.

Additionally, the CF-CAC technique requires other subjects, who only provide anchoring terms to describe their preferred categorization contexts but do not necessarily categorize the whole CiteSeer corpus, to serve as possible neighbors. We develop a questionnaire to support this categorization context collection task. We solicit 68 subjects to participate in our categorization context collection task. Among the 68 subjects, the maximum, minimum, and average numbers of anchoring terms provided are 21, 3, and 7.99 accordingly. We further estimate the intersubject agreement of the anchoring terms between a target subject and all possible neighbors (i.e., the remaining 32 subjects with complete clustering of the CiteSeer corpus and the 68 subjects who only provide anchoring terms) using Jaccard and Dice similarities. The average Jaccard and Dice similarities among the 33 subjects and their possible neighbors are 7.52% and 12.45%, respectively.

4.2 Evaluation Criteria and Procedure

We employ cluster recall and cluster precision (Roussinov & Chen 1999) to measure the effectiveness of the CF-CAC technique and its benchmark technique. To examine the effects of different sizes of anchoring terms on the clustering effectiveness, we randomly sample 80%, 60%, 40% and 20% of anchoring terms from the complete set of anchoring terms of each of the 33 subjects and then investigate their clustering performances using both CF-CAC and CAC techniques. To obtain more reliable tuning results, the described anchoring term sampling and clustering process is performed 5 times, and the overall effectiveness is estimated by averaging the performance estimates obtained from the 5 individual sampling-and-clustering processes.

4.3 Parameter Tuning

In the tuning experiments, we randomly choose the categorization contexts (i.e., anchoring terms) from ten subjects to determine appropriate values for parameters involved in the CAC and CF-CAC techniques. The overall clustering effectiveness of each technique is calculated by averaging the cluster recall and cluster precision obtained from the ten subjects.

We first examine the effects of δ_{DF} (the threshold to remove infrequent features in the feature extraction and selection phase) and α (the threshold to determine whether a feature will be expanded in the anchoring expansion phase) on the effectiveness of the CAC technique. Particularly, we investigate the range of α from 1 to 10 in increments of 0.5. As Figure 2 shows (only a subset of values for α are presented), the best clustering effectiveness of the CAC technique is achieved when α is equal to 2.5. We then tune the value of δ_{DF} from 3 to 10 in increments of 1. As we illustrate in Figure 3 (only a subset of values for δ_{DF} are shown), the CAC technique attains its best performance when δ_{DF} equals to 10. Accordingly, we set α as 2.5 and δ_{DF} as 10 for the subsequent experiments.

The CF-CAC technique involves several parameters, including λ (to remove insignificant relevance weights between anchoring terms) and n (the size of the neighborhood for u_a) in the collaborative context expansion phase, δ_{DF} (to remove infrequent features) in feature extraction and selection phase, and α (to determine whether a feature will be expanded) in the anchoring term expansion phase. Because α and δ_{DF} are also involved in the CAC technique, we choose not to re-conduct the tuning experiments on these two parameters and take the tuning results obtained previously (i.e., $\alpha = 2.5$ and $\delta_{DF} = 10$). In addition, because the purpose of λ is similar to that of α, we also adopt 2.5 for λ. As a result, only the effects of n on the effectiveness of the CF-CAC technique are examined. We range n

from 5 to 20 in increment of 5 and only perform the tuning experiments on n in the case 20% of anchoring terms are used. Our tuning results suggest that the effects of n are marginal. Because CF-CAC with n = 15 slightly outperforms other values for n, we use n = 15 in the subsequent experiments.

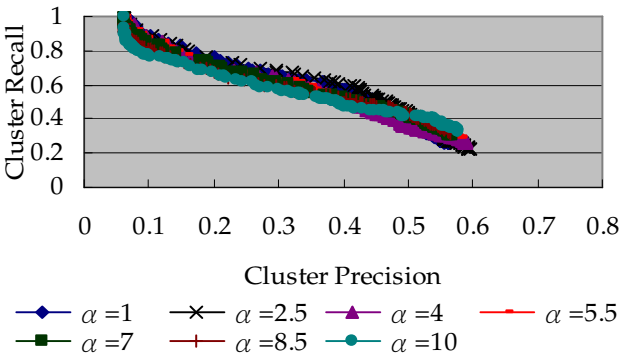

Figure 2: Effects of α for the CAC Technique (Using δ_{DF} = 3)

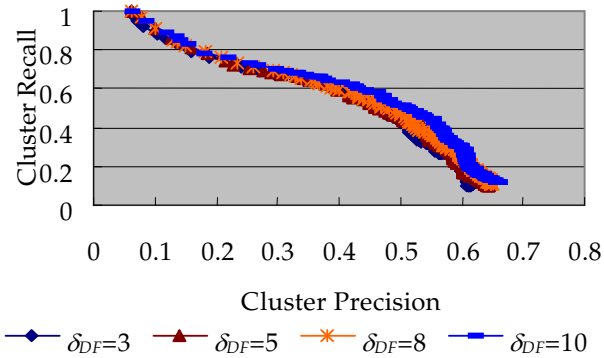

Figure 3: Effects of δ_{DF} for the CAC Technique (Using α = 2.5)

4.4 Comparative Evaluation

Using the parameter values determined previously, we evaluate the effectiveness of the proposed CF-CAC technique and its benchmark technique (i.e., CAC). As we illustrates in Figure 4, when the size of anchoring terms decreases from 100% to 20%, the proposed CF-CAC technique is not sensitive to the sizes of anchoring terms and generally produces comparable clustering results. This is significantly different from that achieved by the CAC technique (shown in Figure 5), which reveals a noticeable sensitivity to the sizes of anchoring terms. We can conclude that the CF-CAC technique is considerably stable over the range of the size of anchoring terms investigated, while the CAC technique is not. To understand the ability of the collaborative context expansion phase of the CF-CAC technique in recovering the discarded anchoring terms of a target user u_a, we calculate the recovery rate that is defined as the percentage of discarded anchoring terms which are reclaimed by the collaborative context expansion phase (i.e., appearing in the expanded categorization context EAT_a). We show the average recovery rate of the 33 subjects in Table 2. Across the range of sizes of anchoring terms input to the CF-CAC technique, the average recovery rate of anchoring terms is greater than 75.74%.

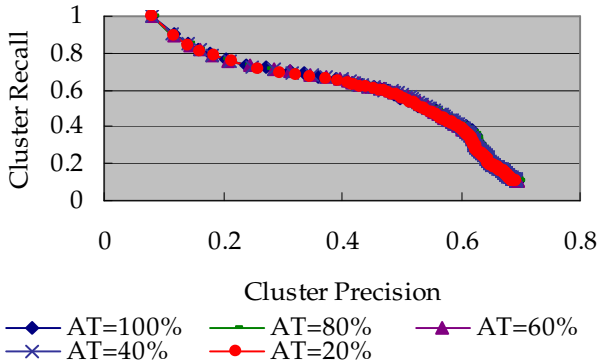

Figure 4: Effects of Sizes of Anchoring Terms for the CF-CAC Technique

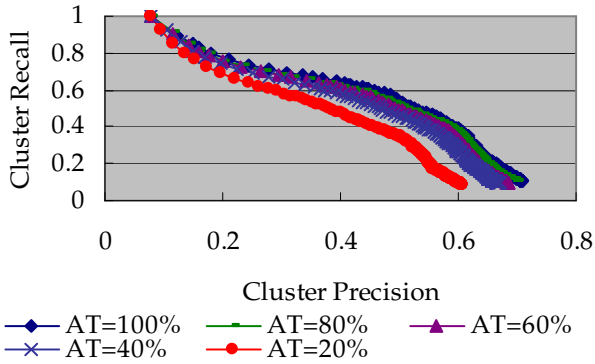

Figure 5: Effects of Sizes of Anchoring Terms for the CAC Technique

Table 2: Recovery Rate of Anchoring Terms of the CF-CAC Technique

	AT = 80%	**AT = 60%**	**AT = 40%**	**AT = 20%**
Recovery Rate	77.81%	77.38%	77.87%	75.74%

We further analyze the comparative performance between the CF-CAC and CAC techniques under different sizes of anchoring terms. Particularly, we calculate the breakeven points (i.e., when cluster recall equal to cluster precision) of the CF-CAC and CAC techniques across different sizes of anchoring terms. As Table 3 shows, the performance differential in breakeven point between the CF-CAC technique and its counterpart increases as the size of anchoring terms decreases. As the size of anchoring terms decreases from 100% to 20%, the effectiveness improvement in breakeven point between CF-CAC and CAC increases from 0.0107 to 0.0917.

Table 3: Effect of Sizes of Anchoring Terms on Breakeven Points of the CF-CAC and CAC Techniques

	AT = 100%	**AT = 80%**	**AT = 60%**	**AT = 40%**	**AT = 20%**
CF-CAC	0.5263	0.5297	0.5306	0.5295	0.5229
CAC	0.5156	0.5078	0.4953	0.4811	0.4312
Improvement*	0.0107	0.0219	0.0353	0.0484	0.0917

*: Improvement = breakeven point of CF-CAC – breakeven point of CAC

5. Conclusion and Future Research Directions

Existing document clustering techniques typically generate a single set of clusters for all individuals without tailoring them to individuals' preferences and contexts and thus are unable to support context-aware document-clustering. Our research is motivated by the importance of and need for context-aware document-clustering. In this study, we extend the CAC technique and propose a Collaborative Filtering-based Context-Aware document-Clustering (CF-CAC) technique by considering not only the target user's but also other users' anchoring terms when approximating the categorization context of the target user. Our empirical evaluation results reveal the superiority, measured by cluster recall and precision, of the CF-CAC technique to the CAC technique.

Some ongoing and future research directions are briefly discussed as follows. First, our evaluation study does not involve a large number of subjects. A future evaluation plan involving more subjects is one of our research directions. Second, our experimental study only includes research articles as our document corpus. Additional empirical evaluation using documents from other domains (e.g., news, patents, etc.) represents an interesting future research direction. Third, the information (specifically, anchoring terms) employed to develop our context-aware document-clustering technique is simply a snapshot at a particular time point. However, users' anchoring terms are usually changed as the time goes by. Therefore, it will be beneficial to incorporate the evolution information of anchoring terms when performing context-aware document-clustering.

Acknowledgment

This work was supported by National Science Council of the Republic of China under the grant NSC 96-2428-H-007-001-MY2.

References

Barreau, D.K. (1991). Context as a factor in personal information management systems. Journal of the American Society for Information Science, 46 (5), 327-339.

Boley, D., Gini, M., Gross, R., Han, E., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, L. (1999). Partitioning-based clustering for web document categorization. Decision Support Systems, 27 (3), 329-341.

Brill, E. (1994). Some advances in rule-based part of speech tagging. In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, WA, 722-727.

Case, D.O. (1991). Conceptual organization and retrieval of text by historians: The role of memory and metaphor. Journal of the American Society for Information Science 42 (9), 657-668.

Cutting, D., Karger, D., Pedersen, J., and Tukey, J. (1992). Scatter/Gather: A cluster-based approach to browsing large document collections. In Proceedings of 15th Annual International ACM SIGIR

Conference, Copenhagen, Denmark, 318-329.
Donovan, J. (1991). Patrons' expectations about collocation: Measuring the difference between psychologically real and the really real. Cataloging and Classification Quarterly, 13 (2), 23-43.
El-Hamdouchi, A. and Willett, P. (1986). Hierarchical document clustering using ward's method. In Proceedings of the 9th Annual International ACM SIGIR Conference, Palazzo dei Congressi, Pisa, Italy, 149-156.
Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30 (11), 964-971.
Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). Data clustering: A review. ACM Computing Surveys, 31 (3), 265-323.
Kwasnik, B.H. (1991). The importance of factors that are not document attributes in the organization of personal documents. Journal of Documentation, 47 (4), 389-398.
Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA, 238-243.
Lakoff, G. (1987). Women, Fire and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press, Chicago.
Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In Proceedings of the 5th ACM SIGKDD International Conference, San Diego, CA, 16-22.
Lin, C., Chen, H., and Nunamaker, J.F. (1999-2000). Verifying the proximity and size hypothesis for self-organizing maps. Journal of Management Information Systems, 16 (3), 57-70.
Mackay, W.E. (1988). Diversity in the use of electronic mail: A preliminary inquiry. ACM Transactions on Office Information Systems, 6 (4), 380-397.
Mackay, W.E. (2000). Responding to cognitive overload: Co-adaptation between users and technology. Intellectica, 30 (1), 177-193.
Pantel, P. and Lin, D. (2002). Document clustering with committees. In Proceedings of the 25th Annual International ACM SIGIR Conference, Tampere, Finland, 199-206.
Quillian, M.R. (1968). Semantic Memory. in Semantic Information Processing, M. Minsky (ed.), The MIT Press, Cambridge, MA, 227-270.
Restorick, F.M. (1986). Novel filing systems applicable to an automated office: A state-of-the-art study. Information Processing and Management, 22 (2), 151-172.
Roussinov, D.G. and Chen, H. (1999). Document clustering for electronic meetings: An experimental comparison of two techniques. Decision Support Systems, 27 (1-2), 67-79.
Rucker, J. and Polanco, M.J. (1997). Siteseer: Personalized navigation for the Web. Communications of the ACM, 40 (3), 73-75.
Turney, P.D. and Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21 (4), 315-346.
Voorhees, E.M. (1986). Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Information Processing and Management, 22 (6), 465-476.
Voutilainen, A. (1993). Nptool: A detector of English noun phrases. In Proceedings of Workshop on Very Large Corpora, Columbus, OH, 48-57.
Wei, C., Chiang, R.H.L., and Wu, C.C. (2006a). Accommodating individual preferences in the categorization of documents: A personalized clustering approach. Journal of Management Information Systems, 23 (2), 173-201.
Wei, C., Hu, P., and Dong, Y.X. (2002). Managing document categories in e-commerce environments: An evolution-based approach. European Journal of Information System, 11 (3), 208-222.
Wei, C, Yang, C.S., Hsiao, H.W., and Cheng, T.H. (2006b). Combining preference- and content-based approaches for improving document clustering effectiveness. Information Processing and Management, 42 (2), 350-372.
Yang, Y. and Chute, C.G. (1994). An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems, 12 (3), 252-277.
Yang, C.S. and Wei, C. (2007). Context-aware document-clustering technique. In Proceedings of 11th Pacific Asia Conference on Information Systems (PACIS 2007), Auckland, New Zealand.